10/865,773 PN-892

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 981 097 A1

# (12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
**23.02.2000 Bulletin 2000/08**

(51) Int. Cl.[7]: **G06F 17/30**

(21) Application number: **98115416.4**

(22) Date of filing: **17.08.1998**

(84) Designated Contracting States:
**AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE**
Designated Extension States:
**AL LT LV MK RO SI**

(71) Applicant:
**Solar Information Co. Ltd.**
**Kaohsiung, Taiwan, R.O.C. (TW)**

(72) Inventor: **Chao, Kuo-Jen**
**Kaohsiung, Taiwan, R.O.C. (TW)**

(74) Representative:
**Bauer, Robert, Dipl.-Ing. et al**
**Patentanwälte Boeters & Bauer**
**Bereiteranger 15**
**81541 München (DE)**

## (54) Search system and method for providing a fulltext search over web pages of world wide web servers

(57) The present invention provides a search system (10) for providing fulltext search over web pages of world wide web servers which can save memory by storing only text, path and hyperlink data of a web page and excluding extraneous data. The system comprises a server (20) connected to an internet (14), a plurality of data groups (22) with web page data, and a management program (24). One user (16) can input search parameter such as keywords into the search system (10) over which the management program (24) uses the search parameters to find matching web pages using an index file (29) within the data groups (22), generates path data for the matched web pages and outputs the path and text data in a standard http format. The search system (10) retrieve only text and path data of each web page and leaves out extraneous data so that the memory space of the server (20) can be saved.

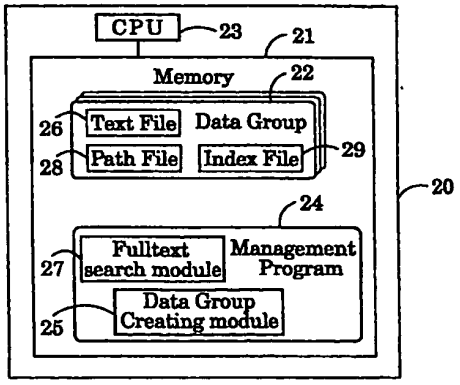


FIG. 2

EP 0 981 097 A1

## Description

[0001] The present invention relates to a search system and method for providing a fulltext search over web pages of world wide web servers.

[0002] The internet has become extremely popular with more and more web servers connecting to it. This enables users to connect with the internet to search a wealth of information. Unfortunately, the vast number of servers currently connected with the internet as well as the number of web pages stored in each server has become unmanageably large thus confusing the user. To overcome this problem, many web page search systems have been produced. Users may key in desired information into the search systems to search servers and web pages.

[0003] To create a database for web pages stored in world wide web servers, the search systems analyze and process data contained in collected web pages from the web servers for use in the search. A single web page may contain many types of files including graphs, text, sound, motion files, etc. Additionally, each web server may contain hundreds, thousands, even tens of thousands, of web pages. Creating a database for even a single web server would be an overwhelming task and the problem is compounded when one considers the fact that a search system must handle hundreds of web servers simultaneously. Clearly, this enormous amount of needed computer memory and increased processing time is unacceptable.

[0004] With this problem in mind, the present invention aims at providing a search system that creates its database by storing the text, path and hyperlink data of a web page only and excluding all extraneous data to solve the above problems.

[0005] This is achieved by the present invention as claimed in claim 1 in that the search system comprises an internet server connected to the internet, a plurality of data groups stored in the server with each of the data groups comprising data from web pages of one world wide web server connected to the internet, and a management program stored in the server for managing operations of the server and providing users with the fulltext search service over the data groups. Each of the data groups in the server comprises a path file for recording path data of each of the web pages in the world wide web server corresponding to the data group and an index file for providing fulltext search for text data contained in the web pages of the world wide web server corresponding to the data group. The management program uses the index file of each data group to find web pages of the corresponding world wide web server which fit the specified search parameter, uses the path file of each data group to find the path data of each of the web pages of the corresponding world wide web server which fit the specified search parameter, and then outputs the result in a predetermined format.

[0006] The invention is illustrated by way of example with reference to the accompanying drawings, in which

Fig.1 is a schematic diagram of a search system for fulltext search web pages of world wide web servers according to the present invention,

Fig.2 is a functional block diagram of the search system shown in Fig.1,

Fig.3 shows a flowchart for creating a database for a web server by the search system shown in Fig.1, and

Fig.4 is a flowchart for fulltext search executed by the search system shown in Fig.1.

[0007] Please refer to Fig.1. Fig.1 is a schematic diagram of the search system 10 for fulltext search of web pages on world wide web servers according to the present invention. Through the internet 14, the search system 10 can connect to the world wide web server 12 and a user 16. The web server 12 usually comprises a home page and a plurality of web pages for the user to search. To create a database, the search system 10 retrieves web page data of the web server 12 and stores only text and path data. This method saves time and memory.

[0008] Please refer to Fig.2. Fig.2 is a functional block diagram of the search system 10 shown in Fig.1. The search system 10 comprises a server 20 connected to the internet 14, a plurality of data groups 22, and a management program 24 stored in the server 20. The server 20 comprises a memory 21 for storing programs and data, and a CPU 23 for executing the program stored in the memory 21. The management program 24 manages the operation of the server 20 and comprises a data group creating module 25 for creating the data groups 22 within the world wide web server 12, and a fulltext search module 27 used by the data groups 22 to perform the fulltext search. Each of the data groups 22 contains data of web pages in a single world wide web server 12, and comprises a text file 26 for recording the text data within the web pages stored in the web server 12, a path file 28 for recording the path of the web pages, and an index file 29 for fulltext search of the text data of the web pages.

[0009] The data group creating module 25 creates the data groups 22 of each web server 12 connected to the internet. The data groups 22 provide fulltext search capability to the user 16. Data groups 22 are made by the data group creating module 25 which first connects to the web server 12 through the internet 14, then uses the text data and path data within each web page to create the text file 26, path file 28, and index file 29. These constitute the data groups 22 of the web server 12.

[0010] The fulltext search module 27 is used for fulltext search of the data groups 22. To search the web pages of the web server 12, the user inputs a keyword or a combination of keywords. Based on this information, the fulltext search module 27 uses the index file 29 to search the text file 26 in each of the data groups 22

for appropriate web pages. Finally, the fulltext search module 27 outputs the text data and path data of the appropriate web pages from the text file 26 and the path file 28 in a standard http web page format. The path file 28 contains the address of the web server 12 and the paths of the web pages whose text data is in the corresponding text file 26.

[0011] Please refer to Fig.3. Fig.3 shows a flowchart for creating the database for the web server 12 by the data group creating module 25 of the search system 10 shown in Fig.1. The flowchart comprises following steps:

step30: connecting to the world wide web server 12 through the internet 14;
step31: creating the text file 26, the path file 28 for the web server 12 and a hyperlink data file, then storing the address of the web server 12 into the path file 28;
step32: requesting the home page of the web server 12;
step33: storing the text data of the home page into the text file 26, storing the path data into the path file 28, storing the hyperlinks into the hyperlink file, creating the index file 29 based on the text data stored in the text file 26, and abandoning all extraneous data in the home page;
step34: using a web page hyperlink from a previously unaccessed hyperlink file to request data from a web page in the web server 12;
step35: storing the text data of the web page into the text file 26, storing the path data into the path file 28, verifying the presence of the hyperlinks not yet stored in the web page and storing them into the hyperlink file, creating the index file 29 based on the text data stored in the text file 26, and then abandoning extraneous data within the web page;
step36: checking if all web pages stored in the hyperlink file are accessed; if not, go back to step 34;
step37: end.

[0012] Using the above procedure, the data group creating module 25 sequentially accesses all web pages in the web server 12 or all or a set number of web pages in a predetermined tree structure, stores text and path data of each web page into the text and path files 26 and 28, respectively, and ignores all extraneous data. This method allows the search system 10 to create data groups 22 efficiently while saving memory space.

[0013] Please refer to Fig.4. Fig.4 is a flowchart showing the fulltext search process by the fulltext search module 27 within the search system 10. The procedure comprises the following steps:

step40: connecting to the search system 10 through the internet 14;
step41: inputting a keyword into the search system 10;
step42: searching the index file 29 of each data group 22 for the corresponding index data based on the keyword;
step43: searching the text file 26 and path file 28 of each data group 22 for corresponding text and path data based on index data corresponding to the keyword;
step44: combining the text and path data, then outputting the data.

[0014] In step 44, the fulltext search module 27, rather than outputting the full text data, outputs the title or a portion of the text data of each web page according to the input command from the user. This output data is arranged in a sequence and format in accordance with the http standard. Since the path data of the searched web pages are stored in each outputted web page in the form of hyperlinks, the user 16 may use hyperlinks to locate the original web server containing the desired web pages.

[0015] When prior art search systems create databases for world wide web servers, the entire web page is often loaded before analyzing and organizing data within the web pages and producing the index data. This process requires a lot of computer memory and processing time. Conversely, the fulltext search system 10 of the present invention saves memory and processing time by storing the text and path data in the web pages of the web server 12 and abandoning extraneous data.

## Claims

1. A search system (10) for providing fulltext search of web pages of world wide web servers connected to an internet (14) comprising:

an internet server (20) connected to the internet (14);
a plurality of data groups (22) stored in the server (20), each of the data groups (22) comprising data from web pages of one world wide web server (12) connected to the internet (14); and
a management program (24) stored in the server (20) for managing operations of the server (20) and providing users with the fulltext search service over the data groups (22);

characterized in that:

each of the data groups (22) in the server (20) comprises:

a path file (28) for recording path data of

each of the web pages in the world wide web server (12) corresponding to the data group (22); and

an index file (29) for providing fulltext search for text data contained in the web pages of the world wide web server (12) corresponding to the data group (22); and

according to at least one user specified search parameter, the management program (24) uses the index file (29) of each data group (22) to find web pages of the corresponding world wide web server (12) which fit the specified search parameter, uses the path file (28) of each data group (22) to find the path data of each of the web pages of the corresponding world wide web server (12) which fit the specified search parameter, and then outputs the result in a predetermined format.

2. The search system (10) of claim 1 wherein each of the data groups (22) stored in the server (20) further comprises a text file (26) for recording the text data contained in each of the web pages of the corresponding world wide web server (12), the path file (28) of each data group (22) is used for recording the path data of each of the web pages contained in the text file (26) of the same data group (22), and the index file (29) of each data group (22) is used for providing fulltext search for the text data contained in the text file (26) of the same data group (22); and wherein after the specified search parameter is provided, the management program (24) uses the index file (29) of each data group (22) to search the text file (26) of the same data group (22) for web pages which fit the search parameter, uses the text file (26) of the same data group (22) to retrieve text data of each web page which fit the search parameter, uses the path file (28) of the same data group (22) to find out the path data of each of the web pages which fit the specified search parameter, and then outputs the result in a predetermined format.

3. The search system (10) of claim 2 wherein the management program (24) outputs the text data and path data of the web pages which fit the specified search parameter in accordance with the http standard web page format.

4. The search system (10) of claim 2 wherein the management program (24) outputs a title portion or part of the text data contained in the web pages which fit the specified search parameter.

5. The search system (10) of claim 2 wherein the search parameter is a keyword or a combination of keywords.

6. The search system (10) of claim 2 wherein the path file (28) of each data group (22) comprises internal paths of all the web pages of the corresponding world wide web server (12) and the internet address of the world wide web server (12) on the internet (14), and wherein the internal paths and the internet address are included in the path data outputted by the management program (24).

7. The search system (10) of claim 2 wherein the management program (24) further comprises a data group creating module (25) for creating the data group (22) of each of the world wide web servers for fulltext search, and wherein when creating one data group (22) for a world wide web server (12), the data group creating module (25) connects the world wide web server (12) through the internet (14) first, retrieves text and path data stored in the web pages of the world wide web server (12) , creates one text file (26) and one path file (28) using the retrieved data, and then creates one index file (29) using the text file (26) for fulltext search of the text data contained in the text file (26).

8. The search system (10) of claim 7 wherein after retrieving the text data and path data contained in each of the web pages, the management program (24) abandons all the other data to save memory space.

9. A method for creating a data group (22) for a world wide web server (12) connected to an internet (14) in a fulltext search system (10), the search system (10) comprising:

an internet server (20) connected to the internet (14) for storing the data group (22) of the world wide web server (12) ; and

a management program (24) stored in the server (20) for managing operations of the server (20) and creating the data group (22) of the world wide web server (12);

the data group (22) of the world wide web server (12) comprising:

a path file (28) for recording path data of each of the web pages in the world wide web server (12); and

an index file (29) for providing fulltext search for text data contained in the web pages in the world wide web server (12);

the method of creating the data group (22) comprising:

connecting the server (20) with the world wide web server (12) through the internet (14);

retrieving path data from each of the web pages of the world wide web server (12) to create the path file (28);

using text data contained in each of the web pages of the world wide web server (12) to create the index file (29) for providing fulltext search over the text data of the web pages in the world wide web server (12).

10. The method of claim 9 wherein the data group (22) of the world wide web server (12) further comprises a text file (26) for recording the text data contained in each of the web pages of the world wide web server (12), the path file (28) of the data group (22) is used for recording the path data of each of the web pages contained in the text file (26) of the data group (22), and the index file (29) of the data group (22) is used for providing fulltext search for the text data contained in the text file (26) of the data group (22) ; and wherein the method further comprises the following step:

retrieving text data from each of the web pages of the world wide web server (12) to create the text file (26).

11. The method of claim 10 wherein after retrieving the text data and path data contained in each of the web pages, the management program (24) abandons all the other data to save memory space.

12. The method of claim 10 wherein when retrieving the text data and path data contained in each of the web pages, the management program (24) can retrieve the data from all the web pages, a predetermined number of web pages, or all of the web pages in a predetermined tree structure from the world wide web server (12) .

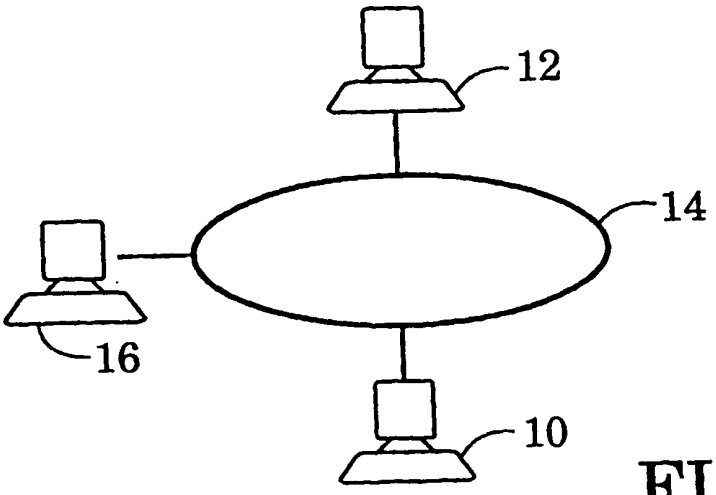
12
14
16
10

FIG. 1

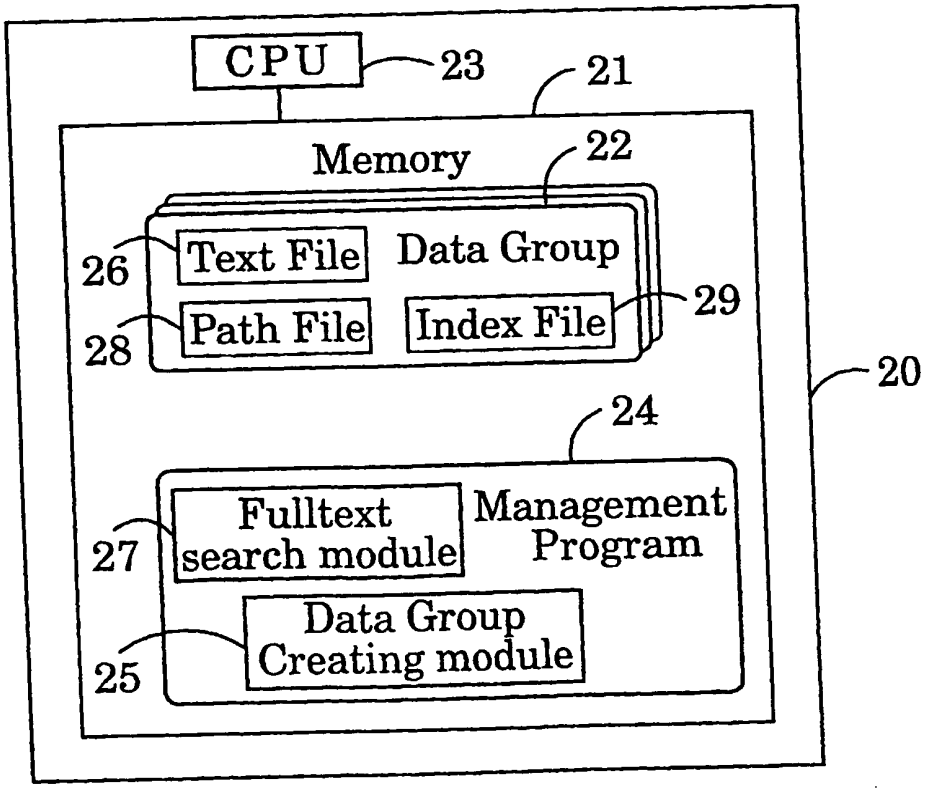
CPU
23
21
Memory
22
26
Text File
Data Group
29
28
Path File
Index File
20
24
27
Fulltext search module
Management Program
Data Group Creating module
25

FIG. 2

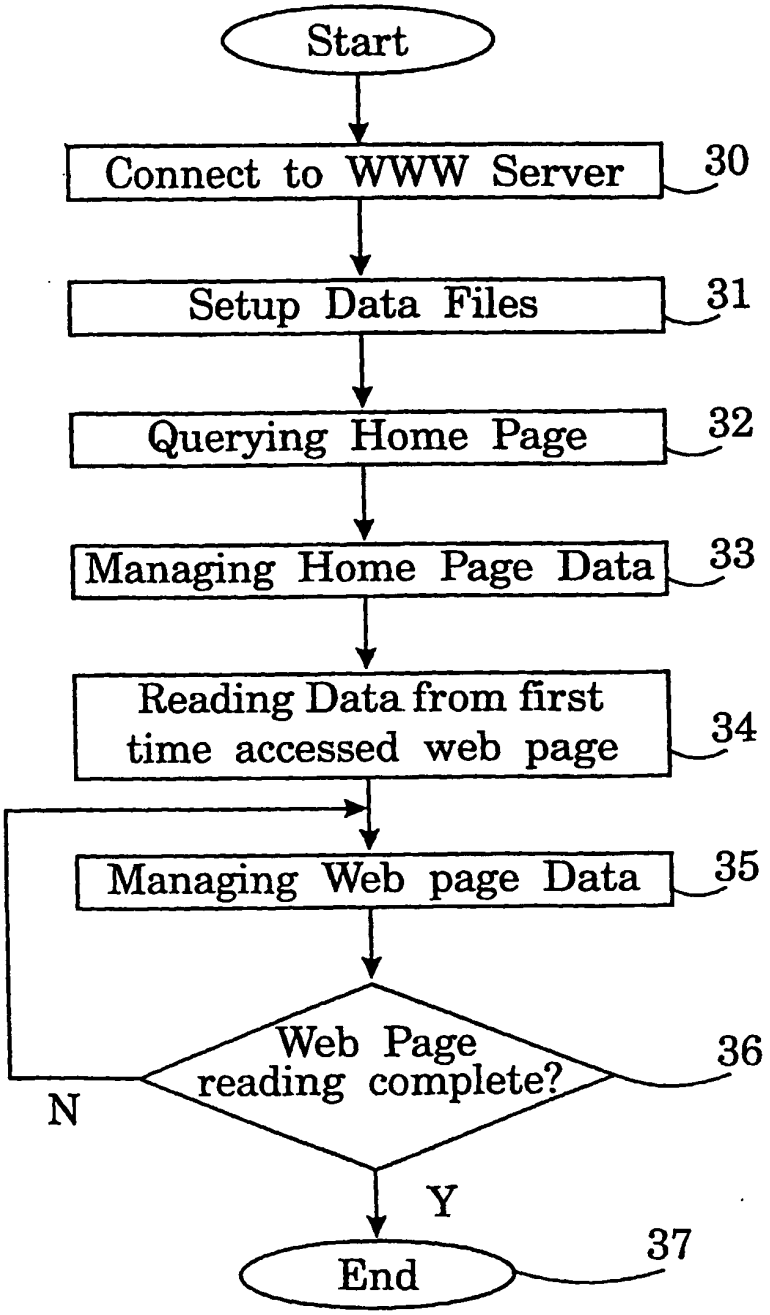
Start
Connect to WWW Server
30
Setup Data Files
31
Querying Home Page
32
Managing Home Page Data
33
Reading Data from first time accessed web page
34
Managing Web page Data
35
Web Page reading complete?
36
N
Y
End
37

FIG. 3

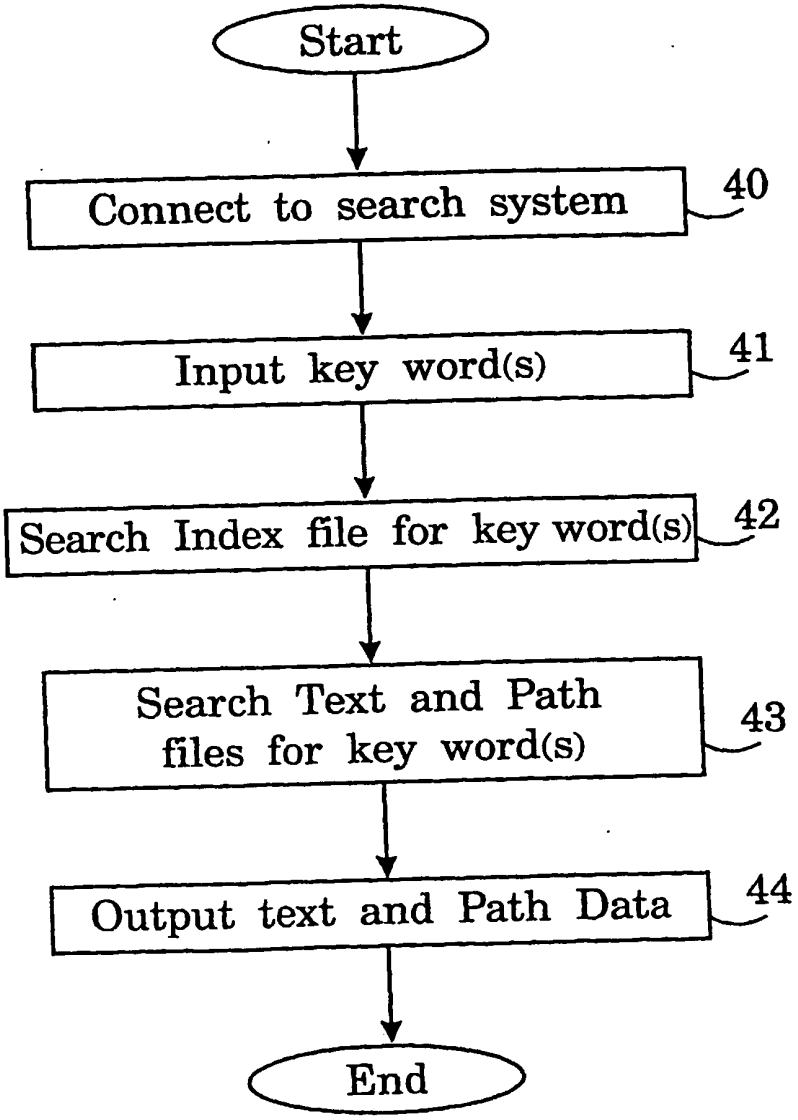
Start
Connect to search system
40
Input key word(s)
41
Search Index file for key word(s)
42
Search Text and Path files for key word(s)
43
Output text and Path Data
44
End

FIG. 4

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 98 11 5416

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
|---|---|---|---|
| Y | LEGH-SMITH J: "NAVIGATING ON-LINE SERVICE ENVIRONMENTS" BRITISH TELECOMMUNICATIONS ENGINEERING, vol. 15, no. 1, 1 April 1996, pages 66-71, XP000589214 * page 67, left-hand column, line 36 - page 68, right-hand column, line 18; figures 1,2 * | 1-6,9, 10,12 | G06F17/30 |
| Y | BRIN S. ET AL.: "The anatomy of a large-scale hypertextual Web search engine" COMPUTER NETWORKS. INTERNATIONAL JOURNAL OF DISTRIBUTION INFORMATIQUE., vol. 30, 1998, pages 107-117, XP002089959 AMSTERDAM NL * page 111, left-hand column, line 33 - page 113, right-hand column, line 21 * | 1-6,9, 10,12 | |
| Y | YUWONO B ET AL: "SEARCH AND RANKING ALGORITHMS FOR LOCATING RESOURCES ON THE WORLD WIDE WEB" PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE ON DATA ENGINEERING, NEW ORLEANS, FEB. 26 - MAR. 1, 1996, no. CONF. 12, 26 February 1996, pages 164-171, XP000632592 SU S Y W (ED ) * page 165, right-hand column, line 30 - page 166, right-hand column, line 43 * | 1-6,9, 10,12 | TECHNICAL FIELDS SEARCHED (Int.Cl.6) G06F |

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| BERLIN | 13 January 1999 | Deane, E |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)